

# Predictive Correlation Screening: Application to Two-stage Predictor Design in High Dimension

Hamed Firouzi, *University of Michigan*, Bala Rajaratnam, *Stanford University*, Alfred O. Hero, *University of Michigan*

## Abstract

We introduce a new approach to variable selection, called Predictive Correlation Screening, for predictor design. Predictive Correlation Screening (PCS) implements false positive control on the selected variables, is well suited to small sample sizes, and is scalable to high dimensions. We establish asymptotic bounds for Familywise Error Rate (FWER), and resultant mean square error of a linear predictor on the selected variables. We apply Predictive Correlation Screening to the following two-stage predictor design problem. An experimenter wants to learn a multivariate predictor of gene expressions based on successive biological samples assayed on mRNA arrays. She assays the whole genome on a few samples and from these assays she selects a small number of variables using Predictive Correlation Screening. To reduce assay cost, she subsequently assays only the selected variables on the remaining samples, to learn the predictor coefficients. We show superiority of Predictive Correlation Screening relative to LASSO and correlation learning (sometimes popularly referred to in the literature as marginal regression or simple thresholding) in terms of performance and computational complexity.

## I. INTRODUCTION

Consider the problem of under-determined multivariate linear regression in which training data  $\{\mathbf{Y}_i, X_{i1}, \dots, X_{ip}\}_{i=1}^n$  is given and a linear estimate of the  $q$ -dimensional response vector  $\mathbf{Y}_i$ ,  $1 \leq i \leq n < p$ , is desired:

$$\mathbf{Y}_i = \mathbf{a}_1 X_{i1} + \dots + \mathbf{a}_p X_{ip} + \epsilon_i, \quad 1 \leq i \leq n, \quad (1)$$

This research was supported in part by AFOSR grant FA9550-13-1-0043.

where  $X_{ij}$  is the  $i$ th sample of regressor variable (covariate)  $X_j$ ,  $\mathbf{Y}_i$  is a vector of response variables, and  $\mathbf{a}_j$  is the  $q$ -dimensional vector of regression coefficients corresponding to  $X_j$ ,  $1 \leq i \leq n, 1 \leq j \leq p$ . There are many applications in which the number  $p$  of regressor variables is larger than the number  $n$  of samples. Such applications arise in text processing of internet documents, gene expression array analysis, combinatorial chemistry, and others (Guyon & Elisseeff, 2003). In this  $p \gg n$  situation training a linear predictor becomes difficult due to rank deficient normal equations, overfitting errors, and high computation complexity. Many penalized regression methods have been proposed to deal with this situation, including: LASSO; elastic net; and group LASSO (Guyon & Elisseeff, 2003; Tibshirani, 1996; Efron et al., 2004; Bühlmann, 2006; Yuan & Lin, 2005; Friedman et al., 2001; Bühlmann & Van De Geer, 2011). These methods perform variable selection by minimizing a penalized mean squared error prediction criterion over all the training data. The main drawback of these methods is their high computation requirements for large  $p$ . In this paper we propose a highly scalable approach to under-determined multivariate regression called Predictive Correlation Screening (PCS).

Like recently introduced correlation screening methods (Hero & Rajaratnam, 2011, 2012) PCS screens for connected variables in a correlation graph. However, unlike these correlation screening methods, PCS screens for connectivity in a bipartite graph between the regressor variables  $\{X_1, \dots, X_p\}$  and the response variables  $\{Y_1, \dots, Y_q\}$ . An edge exists in the bipartite graph between regressor variable  $j$  and response variable  $k$  if the thresholded min-norm regression coefficient matrix  $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_p]$  has a non-zero  $kj$  entry. When the  $j$ -th column of this thresholded matrix is identically zero the  $j$ -th regressor variable is thrown out.

PCS differs from correlation learning, also called marginal regression, simple thresholding, and sure independence screening (Genovese et al., 2012; Fan & Lv, 2008), wherein the simple sample cross-correlation matrix between the response variables and the regressor variables is thresholded. Correlation learning does not account for the correlation between regressor variables, which enters into PCS through the pseudo-inverse correlation matrix - a quantity that introduces little additional computational complexity for small  $n$ .

To illustrate our method of PCS we apply it to a two-stage sequential design problem that is relevant to applications where the cost of samples increases with  $p$ . This is true, for example, with gene microarray experiments: a high throughput “full genome” gene chip with  $p = 40,000$  gene probes can be significantly more costly than a smaller assay that tests fewer than  $p = 15,000$  gene probes (see Fig. 1). In this situation a sensible cost-effective approach would be to use a two-stage procedure: first select a smaller number of variables on a few expensive high throughput samples and then construct the predictor on additional

cheaper low throughput samples. The cheaper samples assay only those variables selected in the first stage.

Specifically, we apply PCS to select variables in the first stage of the two-stage procedure. While bearing some similarities, our two-stage PCS approach differs from the many multi-stage adaptive support recovery methods that have been collectively called distilled sensing (Haupt et al., 2011) in the compressive sensing literature. Like two-stage PCS, distilled sensing (DS) performs initial stage thresholding in order to reduce the number of measured variables in the second stage. However, in distilled sensing the objective is to recover a few variables with high mean amplitudes from a larger set of initially measured regressor variables. In contrast, two-stage PCS seeks to recover a few variables that are strongly predictive of a response variable from a large number of initially measured regressor variables and response variables. Furthermore, unlike in DS, in two-stage PCS the final predictor uses all the information on selected variables collected during both stages.

We establish the following theoretical results on PCS and on the two-stage application of PCS. First, we establish Poisson-like limit theorem for the number of variables that pass the PCS screen. This gives a Poisson approximation to the probability of false discoveries that is accurate for small  $n$  and large  $p$ . The Poisson-like limit theorem also specifies a phase transition threshold for the false discovery probability. Second, with  $n$ , the number of samples in the first stage, and  $t$ , the total number of samples, we establish that  $n$  needs only be of order  $\log(p)$  for two-stage PCS to succeed with high probability in recovering the support set of the optimal OLS predictor. Third, given a cost-per-sample that is linear in the number of assayed variables, we show that the optimal value of  $n$  is on the order of  $\log(t)$ . These three results are analogous to theory for correlation screening (Hero & Rajaratnam, 2011, 2012), support recovery for multivariate lasso (Obozinski et al., 2008), and optimal exploration vs exploitation allocation in multi-armed bandits (Audibert et al., 2007).

The paper is organized as follows. Section II defines the under-determined multivariate regression problem. Section III gives the Poisson-like asymptotic theorem for the thresholded regression coefficient matrix. Section IV defines the PCS procedure and associated p-values. Section V defines the two-stage PCS and prediction algorithm. Section VI gives theorems on support recovery and optimal sample allocation to the first stage of the two-stage algorithm. Section VII presents simulation results and an application to symptom prediction from gene expression data.

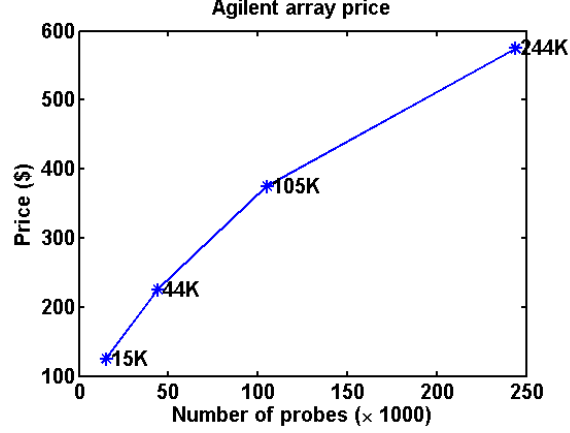


Fig. 1. Pricing per slide for Agilent Custom Micorarrays G2309F, G2513F, G4503A, G4502A (Feb 2013). The cost increases as a function of probeset size. Source: BMC Genomics and RNA Profiling Core.

## II. UNDER-DETERMINED MULTIVARIATE REGRESSION PROBLEM

Assume  $\mathbf{X} = [X_1, \dots, X_p]$  and  $\mathbf{Y} = [Y_1, \dots, Y_q]$  are random vectors of regressor and response variables, from which  $n$  observations are available. We represent the  $n \times p$  and  $n \times q$  data matrices as  $\mathbb{X}$  and  $\mathbb{Y}$ , respectively. We assume that the vector  $\mathbf{X}$  has an elliptically contoured density with mean  $\mu_x$  and non-singular  $p \times p$  covariance matrix  $\Sigma_x$ , i.e. the probability density function is of the form  $f_{\mathbf{X}}(\mathbf{x}) = g((\mathbf{x} - \mu_x)^T \Sigma_x^{-1} (\mathbf{x} - \mu_x))$ , in which  $g$  is a non-negative integrable function. Similarly, the vector  $\mathbf{Y}$ , is assumed to follow an elliptically contoured density with mean  $\mu_y$  and non-singular  $q \times q$  covariance matrix  $\Sigma_y$ . We assume that the joint density function of  $\mathbf{X}$  and  $\mathbf{Y}$  is bounded and differentiable. Denote the  $p \times q$  population cross covariance matrix between  $\mathbf{X}$  and  $\mathbf{Y}$  by  $\Sigma_{xy}$ .

The  $p \times p$  sample covariance matrix  $\mathbf{S}$  for data  $\mathbb{X}$  is defined as:

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{X}_{(i)} - \bar{\mathbf{X}})^T (\mathbf{X}_{(i)} - \bar{\mathbf{X}}), \quad (2)$$

where  $\mathbf{X}_{(i)}$  is the  $i$ th row of data matrix  $\mathbb{X}$ , and  $\bar{\mathbf{X}}$  is the vector average of all  $n$  rows of  $\mathbb{X}$ .

Consider the  $n \times (p+q)$  concatenated matrix  $\mathbb{Z} = [\mathbb{X}, \mathbb{Y}]$ . The sample cross covariance matrix  $\mathbf{S}^{yx}$  is defined as the lower left  $q \times p$  block of the  $(p+q) \times (p+q)$  sample covariance matrix obtained by (2) using  $\mathbb{Z}$  as the data matrix instead of  $\mathbb{X}$ .

Assume that  $p \gg n$ . We define the ordinary least squares (OLS) estimator of  $\mathbf{Y}$  given  $\mathbf{X}$  as the min-norm solution of the underdetermined least squares regression problem

$$\min_{\mathbf{B}} \|\mathbb{Y}^T - \mathbf{B}\mathbb{X}^T\|_F^2, \quad (3)$$

where  $\|\mathbf{A}\|_F$  represents the Frobenius norm of matrix  $\mathbf{A}$ . The min-norm solution to (3) is the  $q \times p$  matrix of regression coefficients

$$\mathbf{B} = \mathbf{S}^{yx}(\mathbf{S}^x)^\dagger, \quad (4)$$

where  $\mathbf{A}^\dagger$  denotes the Moore-Penrose pseudo-inverse of matrix  $\mathbf{A}$ . If the  $i$ th column of  $\mathbf{B}$  is zero then the  $i$ th variable is not included in the OLS estimator. This is the main motivation for the proposed partial correlation screening procedure.

The PCS procedure for variable selection is based on the U-score representation of the correlation matrices. It is easily shown that there exist matrices  $\mathbb{U}^x$  and  $\mathbb{U}^y$  of dimensions  $(n-1) \times p$  and  $(n-1) \times q$  respectively, such that the columns of  $\mathbb{U}^x$  and  $\mathbb{U}^y$  lie on the  $(n-2)$ -dimensional unit sphere  $S_{n-2}$  in  $\mathbb{R}^{n-1}$  and the following representations hold (Hero & Rajaratnam, 2012):

$$\mathbf{S}^{yx} = \mathbf{D}_{\mathbf{S}^y}^{\frac{1}{2}} ((\mathbb{U}^y)^T \mathbb{U}^x) \mathbf{D}_{\mathbf{S}^x}^{\frac{1}{2}}, \quad (5)$$

and:

$$(\mathbf{S}^x)^\dagger = \mathbf{D}_{\mathbf{S}^x}^{-\frac{1}{2}} ((\mathbb{U}^x)^T (\mathbb{U}^x (\mathbb{U}^x)^T)^{-2} \mathbb{U}^x) \mathbf{D}_{\mathbf{S}^x}^{-\frac{1}{2}}, \quad (6)$$

where  $\mathbf{D}_{\mathbf{M}}$  denotes the diagonal matrix obtained by zeroing out the off-diagonals of matrix  $\mathbf{M}$ . Note that  $\mathbb{U}^x$  and  $\mathbb{U}^y$  are constructed from data matrices  $\mathbb{X}$  and  $\mathbb{Y}$ , respectively.

Throughout this paper, we assume the data matrices  $\mathbb{X}$  and  $\mathbb{Y}$  have been normalized in such a way that the sample variance of each variable  $X_i$  and  $Y_j$  is equal to 1 for  $1 \leq i \leq p$  and  $1 \leq j \leq q$ . This simplifies the representations (5) and (6) to  $\mathbf{S}^{yx} = (\mathbb{U}^y)^T \mathbb{U}^x$  and  $(\mathbf{S}^x)^\dagger = (\mathbb{U}^x)^T (\mathbb{U}^x (\mathbb{U}^x)^T)^{-2} \mathbb{U}^x$ . Using these representations, one can write:

$$\hat{\mathbf{Y}} = \mathbf{S}^{yx}(\mathbf{S}^x)^\dagger \mathbf{X} = (\mathbb{U}^y)^T (\mathbb{U}^x (\mathbb{U}^x)^T)^{-1} \mathbb{U}^x \mathbf{X}. \quad (7)$$

Defining  $\tilde{\mathbb{U}}^x = (\mathbb{U}^x (\mathbb{U}^x)^T)^{-1} \mathbb{U}^x \mathbf{D}_{(\mathbb{U}^x)^T (\mathbb{U}^x (\mathbb{U}^x)^T)^{-2} \mathbb{U}^x}^{-\frac{1}{2}}$ , we have:

$$\hat{\mathbf{Y}} = (\mathbb{U}^y)^T \tilde{\mathbb{U}}^x \mathbf{D}_{(\mathbb{U}^x)^T (\mathbb{U}^x (\mathbb{U}^x)^T)^{-2} \mathbb{U}^x}^{\frac{1}{2}} \mathbf{X} \quad (8)$$

$$= (\mathbf{H}^{xy})^T \mathbf{D}_{(\mathbb{U}^x)^T (\mathbb{U}^x (\mathbb{U}^x)^T)^{-2} \mathbb{U}^x}^{\frac{1}{2}} \mathbf{X}, \quad (9)$$

where

$$\mathbf{H}^{xy} = (\tilde{\mathbb{U}}^x)^T \mathbb{U}^y. \quad (10)$$

Note that the columns of matrix  $\tilde{\mathbb{U}}^x$  lie on  $S_{n-2}$ . This can simply be verified by the fact that diagonal entries of the  $p \times p$  matrix  $(\tilde{\mathbb{U}}^x)^T \tilde{\mathbb{U}}^x$  are equal to one.

The U-score representations of covariance matrices completely specify the regression coefficient matrix  $\mathbf{S}^{yx}(\mathbf{S}^x)^\dagger$ .

We define variable selection by discovering columns of the matrix (11) that are not close to zero. The expected number of discoveries will play an important role in the theory of false discoveries, discussed below.

From Sec. II we obtain a U-score representation of the regression coefficient matrix:

$$\mathbf{S}^{yx}(\mathbf{S}^x)^\dagger = (\mathbf{H}^{xy})^T \mathbf{D}_{(\mathbb{U}^x)^T(\mathbb{U}^x(\mathbb{U}^x)^T)^{-2}\mathbb{U}^x}^{\frac{1}{2}}. \quad (11)$$

Under the condition that  $\mathbf{D}_{(\mathbb{U}^x)^T(\mathbb{U}^x(\mathbb{U}^x)^T)^{-2}\mathbb{U}^x}$  has non-zero diagonal entries, the  $i$ th column of  $\mathbf{S}^{yx}(\mathbf{S}^x)^\dagger$  is a zero vector if and only if the  $i$ th row of  $\mathbf{H}^{xy}$  is a zero vector, for  $1 \leq i \leq p$ . This motivates screening for non-zero rows of the matrix  $\mathbf{H}^{xy}$  instead of columns of  $\mathbf{S}^{yx}(\mathbf{S}^x)^\dagger$ .

Fix an integer  $\delta \in \{1, 2, \dots, p\}$  and a real number  $\rho \in [0, 1]$ . For each  $1 \leq i \leq p$ , we call  $i$  a discovery at degree threshold  $\delta$  and correlation threshold  $\rho$  if there are at least  $\delta$  entries in  $i$ th row of  $\mathbf{H}^{xy}$  of magnitude at least  $\rho$ . Note that this definition can be generalized to an arbitrary matrix of the form  $(\mathbb{U}^x)^T \mathbb{U}^y$  where  $\mathbb{U}^x$  and  $\mathbb{U}^y$  are matrices whose columns lie on  $S_{n-2}$ . For a general matrix of the form  $(\mathbb{U}^x)^T \mathbb{U}^y$  we represent the number of discoveries at degree level  $\delta$  and threshold level  $\rho$  as  $N_{\delta, \rho}^{xy}$ .

### III. ASYMPTOTIC THEORY

The following notations are necessary for the propositions in this section. We denote the surface area of the  $(n-2)$ -dimensional unit sphere  $S_{n-2}$  in  $\mathbb{R}^{n-1}$  by  $a_n$ . Assume that  $\mathbf{U}, \mathbf{V}$  are two independent and uniformly distributed random vectors on  $S_{n-2}$ . For a threshold  $\rho \in [0, 1]$ , let  $r = \sqrt{2(1-\rho)}$ .  $P_0$  is then defined as the probability that either  $\|\mathbf{U} - \mathbf{V}\|_2 \leq r$  or  $\|\mathbf{U} + \mathbf{V}\|_2 \leq r$ .  $P_0$  can be computed using the formula for the area of spherical caps on  $S_{n-2}$  (Hero & Rajaratnam, 2012).

Define the index set  $\mathcal{C}$  as:

$$\begin{aligned} \mathcal{C} = \{ & (i_0, i_1, \dots, i_\delta) : \\ & 1 \leq i_0 \leq p, 1 \leq i_1 < \dots < i_\delta \leq q \}. \end{aligned} \quad (12)$$

For arbitrary joint density  $f_{\mathbf{U}_0, \dots, \mathbf{U}_\delta}(\mathbf{u}_0, \dots, \mathbf{u}_\delta)$  defined on the Cartesian product  $S_{n-2}^{\delta+1} = S_{n-2} \times \dots \times S_{n-2}$ , define  $\overline{f_{\mathbf{U}_0^x, \mathbf{U}_{i_1}^y, \dots, \mathbf{U}_{i_\delta}^y}}(\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_\delta)$  as the average of

$$\begin{aligned} & f_{\mathbf{U}_i^x}(s_0 \mathbf{u}_0, s_1 \mathbf{u}_1, \dots, s_\delta \mathbf{u}_\delta) = \\ & f_{\mathbf{U}_{i_0}^x, \mathbf{U}_{i_1}^y, \dots, \mathbf{U}_{i_\delta}^y}(s_0 \mathbf{u}_0, s_1 \mathbf{u}_1, \dots, s_\delta \mathbf{u}_\delta), \end{aligned} \quad (13)$$

for all  $\vec{i} = (i_0, i_1, \dots, i_\delta) \in \mathcal{C}$  and  $s_j \in \{-1, 1\}, 0 \leq j \leq \delta$ .

In the following propositions,  $k$  represents an upper bound on the number of non-zero entries in any row or column of covariance matrix  $\Sigma_x$  or cross covariance matrix  $\Sigma_{xy}$ . We define  $\|\Delta_{p,q,n,k,\delta}^{xy}\|_1 = |\mathcal{C}|^{-1} \sum_{\vec{i} \in \mathcal{C}} \Delta_{p,q,n,k,\delta}^{xy}(\vec{i})$ , the average dependency coefficient, as the average of

$$\Delta_{p,q,n,k,\delta}^{xy}(\vec{i}) = \left\| (f_{\mathbf{U}_{\vec{i}}|\mathbf{U}_{A_k(i_0)}} - f_{\mathbf{U}_{\vec{i}}}) / f_{\mathbf{U}_{\vec{i}}} \right\|_\infty, \quad (14)$$

in which  $A_k(i_0)$  is defined as the set complement of the union of indices of non-zero elements of the  $i_0$ -th column of  $\Sigma_{yx}\Sigma_x^{-1}$ . Finally, the function  $J$  of the joint density  $f_{\mathbf{U}_0, \dots, \mathbf{U}_\delta}(\mathbf{u}_0, \dots, \mathbf{u}_\delta)$  is defined as:

$$J(f_{\mathbf{U}_0, \dots, \mathbf{U}_\delta}) = |S_{n-2}|^\delta \int_{S_{n-2}} f_{\mathbf{U}_0, \dots, \mathbf{U}_\delta}(\mathbf{u}, \dots, \mathbf{u}) d\mathbf{u}. \quad (15)$$

The following proposition gives an asymptotic expression for the number of discoveries in a matrix of the form  $(\mathbf{U}^x)^T \mathbf{U}^y$ , as  $p \rightarrow \infty$ , for fixed  $n$ . Also it states that, under certain assumptions, the probability of having at least one discovery converges to a given limit. This limit is equal to the probability that a certain Poisson random variable  $N_{\delta, \rho_p}^*$  with rate equal to  $\lim_{p \rightarrow \infty} E[N_{\delta, \rho_p}^{xy}]$  takes a non-zero value, i.e. it satisfies:  $N_{\delta, \rho_p}^* > 0$ .

*Proposition 1:* Let  $\mathbf{U}^x = [\mathbf{U}_1^x, \mathbf{U}_2^x, \dots, \mathbf{U}_p^x]$  and  $\mathbf{U}^y = [\mathbf{U}_1^y, \mathbf{U}_2^y, \dots, \mathbf{U}_q^y]$  be  $(n-1) \times p$  and  $(n-1) \times q$  random matrices respectively, with  $\mathbf{U}_i^x, \mathbf{U}_j^y \in S_{n-2}$  for  $1 \leq i \leq p, 1 \leq j \leq q$ . Fix integers  $\delta \geq 1$  and  $n > 2$ . Assume that the joint density of any subset of  $\{\mathbf{U}_1^x, \dots, \mathbf{U}_p^x, \mathbf{U}_1^y, \dots, \mathbf{U}_q^y\}$  is bounded and differentiable. Let  $\{\rho_p\}_p$  be a sequence in  $[0, 1]$  such that  $\rho_p \rightarrow 1$  as  $p \rightarrow \infty$  and  $p^{\frac{1}{\delta}} q (1 - \rho_p^2)^{\frac{(n-2)}{2}} \rightarrow e_{n,\delta}$ . Then,

$$\begin{aligned} \lim_{p \rightarrow \infty} E[N_{\delta, \rho_p}^{xy}] &= \lim_{p \rightarrow \infty} \xi_{p,q,n,\delta,\rho_p} J(\overline{f_{\mathbf{U}_*^x, \mathbf{U}_{\bullet 1}^y, \dots, \mathbf{U}_{\bullet \delta}^y}}) \\ &= \kappa_{n,\delta} \lim_{p \rightarrow \infty} J(\overline{f_{\mathbf{U}_*^x, \mathbf{U}_{\bullet 1}^y, \dots, \mathbf{U}_{\bullet \delta}^y}}), \end{aligned} \quad (16)$$

where  $\xi_{p,q,n,\delta,\rho_p} = p \binom{q}{\delta} P_0^\delta$  and  $\kappa_{n,\delta} = (e_{n,\delta} a_n / (n-2))^\delta / \delta!$ .

Assume also that  $k = o((p^{\frac{1}{\delta}} q)^{1/(\delta+1)})$  and that the average dependency coefficient satisfies

$\lim_{p \rightarrow \infty} \|\Delta_{p,q,n,k,\delta}^{xy}\|_1 = 0$ . Then:

$$p(N_{\delta, \rho_p}^{xy} > 0) \rightarrow 1 - \exp(-\Lambda_\delta^{xy}), \quad (17)$$

with

$$\Lambda_\delta^{xy} = \lim_{p \rightarrow \infty} E[N_{\delta, \rho_p}^{xy}]. \quad (18)$$

*Proof of Proposition 1:* See appendix.

The following proposition states that when the rows of data matrices  $\mathbb{X}$  and  $\mathbb{Y}$  are i.i.d. elliptically distributed with block sparse covariance matrices, the rate (16) in Proposition 1 becomes independent of  $\Sigma_x$  and  $\Sigma_{xy}$ . Specifically, the  $(\delta + 1)$ -fold average  $J(\overline{f_{\mathbf{U}_\bullet^x, \mathbf{U}_{*1}^y, \dots, \mathbf{U}_{*\delta}^y}})$  converges to 1 while the average dependency coefficient  $\|\Delta_{p,q,n,k,\delta}^{xy}\|_1$  goes to 0, as  $p \rightarrow \infty$ . This proposition will play an important role in identifying phase transitions and in approximating  $p$ -values.

*Proposition 2:* Assume the hypotheses of Prop. 1 are satisfied. In addition assume that the rows of data matrices  $\mathbb{X}$  and  $\mathbb{Y}$  are i.i.d. elliptically distributed with block sparse covariance and cross covariance matrices  $\Sigma_x$  and  $\Sigma_{xy}$ . Then  $\Lambda_\delta^{xy}$  in the limit (18) in Prop. 1 is equal to the constant  $\kappa_{n,\delta}$  given in (16). Moreover,  $\tilde{\mathbb{U}}_x \approx \mathbb{U}_x$ .

*Proof of Proposition 2:* See appendix.

#### IV. PREDICTIVE CORRELATION SCREENING

Under the assumptions of Propositions 1 and 2:

$$p(N_{\delta,\rho_p}^{xy} > 0) \rightarrow 1 - \exp(-\xi_{p,q,n,\delta,\rho_p}) \quad \text{as } p \rightarrow \infty \quad (19)$$

Using the above limit, approximate  $p$ -values can be computed. Fix a degree threshold  $\delta \leq q$  and a correlation threshold  $\rho^* \in [0, 1]$ . Define  $\mathcal{G}_{\rho^*}(\mathbf{H}^{xy})$  as the undirected bipartite graph (Fig. 2) with parts labeled  $x$  and  $y$ , vertices  $\{X_1, X_2, \dots, X_p\}$  in part  $x$  and  $\{Y_1, Y_2, \dots, Y_q\}$  in part  $y$ . For  $1 \leq i \leq p$  and  $1 \leq j \leq q$ , vertices  $X_i$  and  $Y_j$  are connected if  $|h_{ij}^{xy}| > \rho^*$ , where  $h_{ij}^{xy}$  is the  $(i, j)$ th entry of  $\mathbf{H}^{xy}$  defined in (10). Denote by  $d_i^x$  the degree of vertex  $X_i$  in  $\mathcal{G}_{\rho^*}(\mathbf{H}^{xy})$ . For each value  $\delta \in \{1, \dots, \max_{1 \leq i \leq p} d_i^x\}$ , and each  $i$ ,  $1 \leq i \leq p$ , denote by  $\rho_i(\delta)$  the maximum value of the correlation threshold  $\rho$  for which  $d_i^x \geq \delta$  in  $\mathcal{G}_\rho(\mathbf{H}^{xy})$ .  $\rho_i(\delta)$  is in fact equal to the  $\delta$ th largest value  $|h_{ij}^{xy}|$ ,  $1 \leq j \leq q$ .  $\rho_i(\delta)$  can be computed using Approximate Nearest Neighbors (ANN) type algorithms (Jégou et al., 2011; Arya et al., 1998). Now for each  $i$  define the modified threshold  $\rho_i^{\text{mod}}(\delta)$  as:

$$\rho_i^{\text{mod}}(\delta) = w_i \rho_i(\delta), \quad 1 \leq i \leq p, \quad (20)$$

where  $w_i = D(i) / \sum_{j=1}^p D(j)$ , in which  $D(i)$  is the  $i$ th diagonal element of the diagonal matrix  $\mathbf{D}_{(\mathbb{U}^x)^T (\mathbb{U}^x (\mathbb{U}^x)^T)^{-2} \mathbb{U}^x}$  (recall Sec. II).

Using Propositions 1 and 2 the  $p$ -value associated with variable  $X_i$  at degree level  $\delta$  can be approximated as:

$$pv_\delta(i) \approx 1 - \exp(-\xi_{p,q,n,\delta,\rho_i^{\text{mod}}(\delta)}). \quad (21)$$



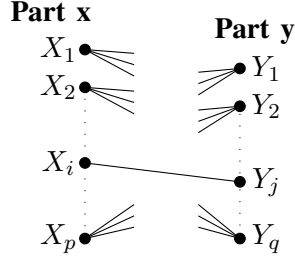


Fig. 2. Predictive correlation screening thresholds the matrix  $\mathbf{H}^{xy}$  in (11) to find variables  $X_i$  that are most predictive of responses  $Y_j$ . This is equivalent to finding sparsity in a bipartite graph  $\mathcal{G}_{\rho^*}(\mathbf{H}^{xy})$  with parts  $x$  and  $y$  which have  $p$  and  $q$  vertices, respectively. For  $1 \leq i \leq p$  and  $1 \leq j \leq q$ , vertex  $X_i$  in part  $x$  is connected to vertex  $Y_j$  in part  $y$  if  $|h_{ij}^{xy}| > \rho^*$ .

The set of p-values (21),  $i = 1, \dots, p$ , provides a measure of importance of each variable  $X_i$  in predicting  $Y_j$ 's. Under a block-sparsity null hypothesis, the most important variables would be the ones that have the smallest p-values. Similar to the result in (Hero & Rajaratnam, 2011, 2012), there is a phase transition in the p-values as a function of threshold  $\rho$ . More exactly, there is a critical threshold  $\rho_{c,\delta}$  such that if  $\rho > \rho_{c,\delta}$ , the average number  $E[N_{\delta,\rho}^{xy}]$  of discoveries abruptly decreases to 0 and if  $\rho < \rho_{c,\delta}$  the average number of discoveries abruptly increases to  $p$ . The value of this critical threshold is:

$$\rho_{c,\delta} = \sqrt{1 - (c_{n,\delta}^{xy} p)^{-2\delta/(\delta(n-2)-2)}}, \quad (22)$$

where  $c_{n,\delta}^{xy} = a_n \delta J(\overline{f_{\mathbf{U}_\bullet^x, \mathbf{U}_{*1}^y, \dots, \mathbf{U}_{*\delta}^y}})$ . When  $\delta = 1$ , the expression given in (22) is identical, except for the constant  $c_{n,\delta}^{xy}$ , to the expression (3.14) in (Hero & Rajaratnam, 2011).

Expression (22) is useful in choosing the PCS correlation threshold  $\rho^*$ . Selecting  $\rho^*$  slightly greater than  $\rho_{c,\delta}$  will prevent the bipartite graph  $\mathcal{G}_{\rho^*}(\mathbf{H}^{xy})$  from having an overwhelming number of edges.

Normally  $\delta = 1$  would be selected to find all regressor variables predictive of at least 1 response variable  $Y_j$ . A value of  $\delta = d > 1$  would be used if the experimenter were only interested in variables that were predictive of at least  $d$  of the responses. Pseudo-code for the complete algorithm for variable selection is shown in Fig. 3. The worse case computational complexity of the PCS algorithm is only  $O(np \log q)$ .

## V. TWO-STAGE PREDICTOR DESIGN

Assume there are a total of  $t$  samples  $\{\mathbf{Y}_i, \mathbf{X}_i\}_{i=1}^t$  available. During the first stage a number  $n \leq t$  of these samples are assayed for all  $p$  variables and during the second stage the rest of the  $t - n$  samples

- Initialization:
  - 1) Choose an initial threshold  $\rho^* > \rho_{c,\delta}$
  - 2) Calculate the degree of each vertex on side  $x$  of the bipartite graph  $\mathcal{G}_{\rho^*}(\mathbf{H}^{xy})$
  - 3) Select a value of  $\delta \in \{1, \dots, \max_{1 \leq i \leq p} d_i^x\}$
- For each  $i = 1, \dots, p$  find  $\rho_i(\delta)$  as the  $\delta$ th greatest element of  $\{|h_{ij}|, 1 \leq j \leq q\}$
- Compute  $\rho_i^{\text{mod}}(\delta)$  using (20)
- Approximate the p-value corresponding to the  $i$ th independent variable  $X_i$  as  $pv_\delta(i) \approx 1 - \exp(-\xi_{p,q,n,\delta,\rho_i^{\text{mod}}(\delta)})$ .
- Screen variables by thresholding the p-values  $pv_\delta(i)$  at desired significance level

Fig. 3. Predictive Correlation Screening (PCS) Algorithm

are assayed for a subset of  $k \leq p$  of the variables. Subsequently, a  $k$ -variable predictor is designed using all  $t$  samples collected during both stages. The first stage of the PCS predictor is implemented by using the PCS algorithm with  $\delta = 1$ .

As this two-stage PCS algorithm uses  $n$  and  $t$  samples in stage 1 and stage 2 respectively, we denote the algorithm above as the  $n|t$  algorithm. Experimental results in Sec. VII show that for  $n \ll p$ , if LASSO or correlation learning is used instead of PCS in stage 1 of the two-stage predictor the performance suffers. An asymptotic analysis (as the total number of samples  $t \rightarrow \infty$ ) of the above two-stage predictor can be performed to obtain optimal sample allocation rules for stage 1 and stage 2. The asymptotic analysis discussed in Sec. VI provides minimum Mean Squared Error (MSE) under the assumption that  $n$ ,  $t$ ,  $p$ , and  $k$  satisfy the budget constraint:

$$np + (t - n)k \leq \mu, \quad (23)$$

where  $\mu$  is the total budget available. The motivation for this condition is to bound the total sampling cost of the experiment.

## VI. OPTIMAL STAGE-WISE SAMPLE ALLOCATION

We first give theoretical upper bounds on the Family-Wise Error Rate (FWER) of performing variable selection using p-values obtained via PCS. Then, using the obtained bound, we compute the asymptotic optimal sample size  $n$  used in the first stage of the two-stage predictor, introduced in the previous section, to minimize the asymptotic expected MSE.

We assume that the response  $\mathbf{Y}$  satisfies the following ground truth model:

$$\mathbf{Y} = \mathbf{a}_{i_1}X_{i_1} + \mathbf{a}_{i_2}X_{i_2} + \cdots + \mathbf{a}_{i_k}X_{i_k} + \mathbf{N}, \quad (24)$$

where  $\pi_0 = \{i_1, \dots, i_k\}$  is a set of distinct indices in  $\{1, \dots, p\}$ ,  $\mathbf{X} = [X_1, X_2, \dots, X_p]$  is the vector of predictors,  $\mathbf{Y}$  is the  $q$ -dimensional response vector, and  $\mathbf{N}$  is a noise vector statistically independent of  $\mathbf{X}$ .  $X_{i_1}, \dots, X_{i_k}$  are called active variables and the remaining  $p-k$  variables are called inactive variables. We assume that the  $p$ -dimensional vector  $\mathbf{X}$  follows a multivariate normal distribution with mean  $\mathbf{0}$  and  $p \times p$  covariance matrix  $\Sigma = [\sigma_{ij}]_{1 \leq i, j \leq p}$ , where  $\Sigma$  has the following block diagonal structure:

$$\sigma_{ij} = \sigma_{ji} = 0, \quad \forall i \in \pi_0, j \in \{1, \dots, p\} \setminus \pi_0. \quad (25)$$

In other words active (respectively inactive) variables are only correlated with the other active (respectively inactive) variables. Also, we assume that  $\mathbf{N}$  follows a multivariate normal distribution with mean  $\mathbf{0}$  and covariance matrix  $\sigma \mathbf{I}_{q \times q}$ .

We use the PCS algorithm of Sec. IV with  $\delta = 1$  to select the  $k$  variables with the smallest p-values. These selected variables will then be used as estimated active variables in the second stage. The following proposition gives an upper bound on the probability of selection error for the PCS algorithm.

*Proposition 3:* If  $n \geq \Theta(\log p)$  then with probability at least  $1 - q/p$ , PCS recovers the exact support  $\pi_0$ .

*Proof of Proposition 3:* See appendix. □

Proposition 3 can be compared to Thm. 1 in (Obozinski et al., 2008) for recovering the support  $\pi_0$  by minimizing a LASSO-type objective function. The constant in  $\Theta(\log p)$  of Prop. 3 is increasing in the dynamic range coefficient

$$\max_{i=1, \dots, q} \frac{|\pi_0|^{-1} \sum_{j \in \pi_0} |b_{ij}|}{\min_{j \in \pi_0} |b_{ij}|} \in [1, \infty), \quad (26)$$

where  $\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_p] = \Sigma^{1/2} \mathbf{A}$ . The worst case (largest constant in  $\Theta(\log p)$ ) occurs when there is high dynamic range in some rows of the  $q \times p$  matrix  $\mathbf{B}$ .

The following proposition states the optimal sample allocation rule for the two-stage predictor, as  $t \rightarrow \infty$ .

*Proposition 4:* The optimal sample allocation rule for the two-stage predictor introduced in Sec. V under the cost condition (23) is

$$n = \begin{cases} O(\log t), & c(p-k) \log t + kt \leq \mu \\ 0, & o.w. \end{cases} \quad (27)$$

*Proof of Proposition 4:* See appendix.  $\square$

Proposition 4 implies that for a generous budget ( $\mu$  large) the optimal first stage sampling allocation is  $\log(t)$ . However, when the budget is tight it is better to skip stage 1 ( $n = 0$ ). Figure 4 illustrates the allocation region as a function of the sparsity coefficient  $\rho = 1 - k/p$ .

## VII. SIMULATION RESULTS

*a) Efficiency of Predictive Correlation Screening.:* We illustrate the performance of the two-stage PCS algorithm and compare to LASSO and correlation learning methods (Tibshirani, 1996; Genovese et al., 2012).

In the first set of simulations we generated an  $n \times p$  data matrix  $\mathbb{X}$  with independent columns, each of which is drawn from a  $p$ -dimensional multivariate normal distribution with identity covariance matrix. The  $q \times p$  coefficient matrix  $\mathbf{A}$  is then generated in such a way that each column of  $\mathbf{A}$  is active with probability 0.1. Each active column of  $\mathbf{A}$  is a random  $q$ -dimensional vector with i.i.d.  $\mathcal{N}(0, 1)$  entries, and each inactive column of  $\mathbf{A}$  is a zero vector. Finally, a synthetic response matrix  $\mathbb{Y}$  is generated by a simple linear model

$$\mathbb{Y}^T = \mathbf{A}\mathbb{X}^T + \mathbb{N}^T, \quad (28)$$

where  $\mathbb{N}$  is  $n \times q$  noise matrix whose entries are i.i.d.  $\mathcal{N}(0, 0.05)$ . The importance of a variable is measured by the value of the  $\ell_2$  norm of the corresponding column of  $\mathbf{A}$ . Note that the linear model in (28) trivially satisfies the block sparsity assumptions on the covariance matrices in Prop. 2.

We implemented LASSO using an active set type algorithm - claimed to be one the fastest methods for solving LASSO (Kim & Park, 2010). We set the number of regressor and response variables to  $p = 200$  and  $q = 20$ , respectively, while the number of samples  $n$  was varied from 4 to 50. Figure 5 shows the average number of mis-selected variables for both methods, as a function of  $n$ . The plot is computed by averaging the results of 400 independent experiments for each value of  $n$ . Figure 6 shows the average run time on a logarithmic scale, as a function of  $n$  (MATLAB version 7.14 running on 2.80GHz CPU). As we see, for low number of samples, PCS has better performance than LASSO and is significantly faster.

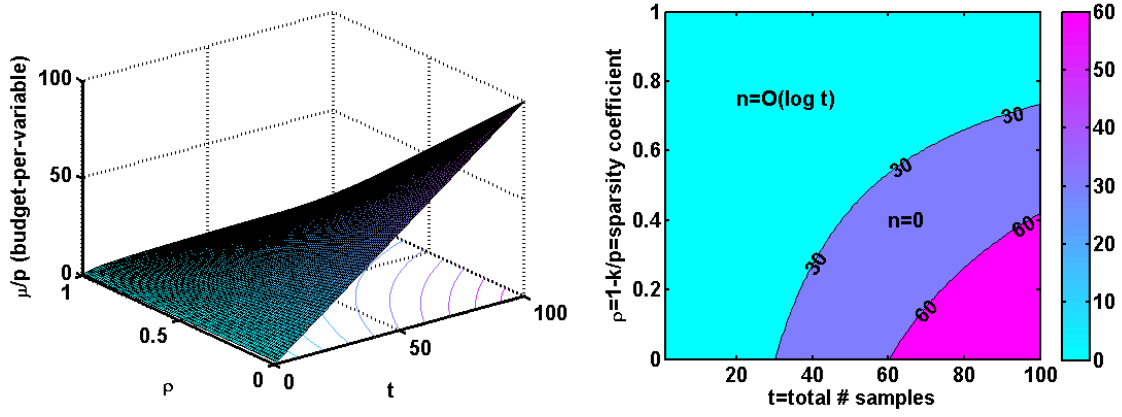


Fig. 4. Left: surface  $\mu/p = \rho \log t + (1-\rho)t$ . Right: contours indicating optimal allocation regions for  $\mu/p = 30$  and  $\mu/p = 60$ . ( $\rho = 1 - k/p$ )

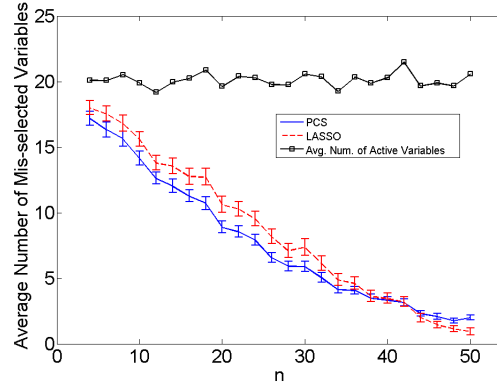


Fig. 5. Average number of mis-selected variables for active set implementation of LASSO (dashed) vs. Predictive Correlation Screening (solid),  $p = 200, q = 20$ .

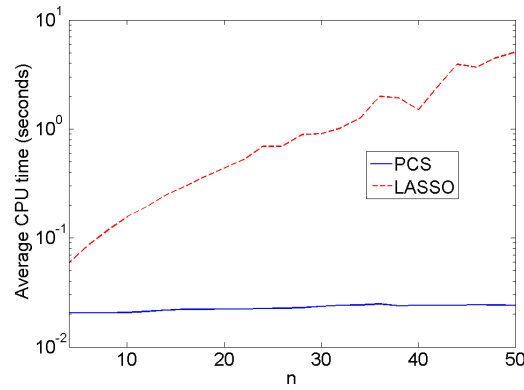


Fig. 6. Average CPU time for active set implementation of LASSO (dashed) vs. PCS (solid),  $p = 200, q = 20$ .

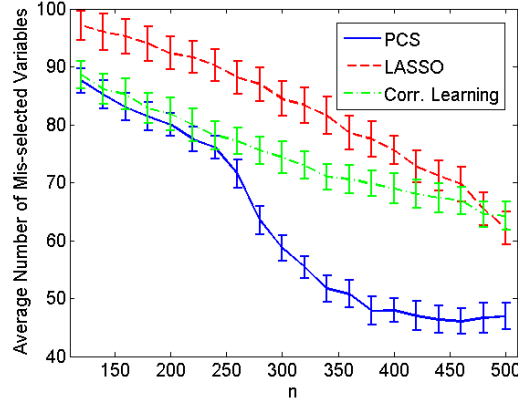


Fig. 7. Average number of mis-selected variables. Active set implementation of LASSO (red-dashed) vs. correlation learning (green-dashed) vs. PCS (solid),  $p = 10^4$ ,  $q = 1$ .

To illustrate PCS for a higher dimensional example, we set  $p = 10^4$ ,  $q = 1$  and compared PCS with LASSO and also the correlation learning method of (Genovese et al., 2012), for a small number of samples. Figure 7 shows the results of this simulation over an average of 400 independent experiments for each value of  $n$ . In this experiment, exactly 100 entries of  $\mathbf{A}$  are active. The active entries are i.i.d. draws of  $\mathcal{N}(0, 1)$  and inactive entries are equal to zero. Unlike Fig. 5, here the regressors variables are correlated. Specifically,  $X_1, \dots, X_p$ , are i.i.d. draws from a multivariate normal distribution with mean  $\mathbf{0}$  and block diagonal covariance matrix satisfying (25). As we see for small number of samples, PCS performs significantly better in selecting the important regressor variables.

*b) Efficiency of The Two-stage Predictor.:* To test the efficiency of the proposed two-stage predictor, a total of  $t$  samples are generated using the linear model (28) from which  $n = 25 \log t$  are used for the task of variable selection at the first stage. All  $t$  samples are then used to compute the OLS estimator restricted to the selected variables. We chose  $t$  such that  $n = (130 : 10 : 200)$ . The performance is evaluated by the empirical  $\text{MSE} := \sum_{i=1}^m (Y_i - \hat{Y}_i)^2 / m$ , where  $m$  is the number of simulation trials. Similar to the previous experiment, exactly 100 entries of  $\mathbf{A}$  are active and the regressor variables follow a multivariate normal distribution with mean  $\mathbf{0}$  and block diagonal covariance matrix of the form (25). Figure 8 shows the result of this simulation for  $p = 10^4$  and  $q = 1$ . Each point on the plot is an average of 100 independent experiments. Observe that in this low sample regime, when LASSO or correlation learning are used instead of PCS in the first stage, the performance suffers.

*c) Estimation of FWER Using Monte Carlo Simulation.:* We set  $p = 1000$ ,  $k = 10$  and  $n = (100 : 100 : 1000)$  and using Monte Carlo simulation, we computed the probability of error (i.e. when the exact

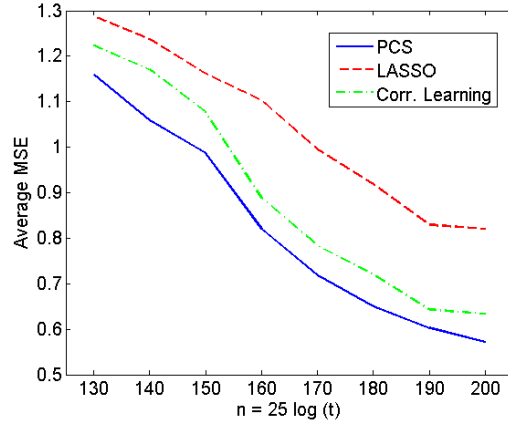


Fig. 8. Prediction MSE for the two-stage predictor when  $n = 25 \log t$  samples are used for screening at the first stage and all  $t$  samples are used for computing the OLS estimator coefficients at the second stage. The solid plot shows the MSE when PCS is used in the first stage while the red and green dashed plots show the MSE when PCS is replaced with LASSO and correlation learning, respectively. Here,  $p = 10^4$  and  $q = 1$ . The Oracle OLS (not shown), which is the OLS predictor constructed on the true support set, has average MSE performance that is a factor of 2 lower than the curves shown in the figure. This is due to the relatively small sample size available to these algorithms.

support is not recovered) for the PCS. In order to prevent the ratios  $|a_j| / \sum_{l \in \pi_0} |a_l|$ ,  $j \in \pi_0$  from getting close to zero, the active coefficients were generated via a Bernoulli-Gaussian distribution of the form:

$$a \sim 0.5\mathcal{N}(1, \sigma^2) + 0.5\mathcal{N}(-1, \sigma^2), \quad (29)$$

Figure 9 shows the estimated probabilities. Each point of the plot is an average of  $N = 10^4$  experiments. As the value of  $\sigma$  decreases dynamic range coefficient (26) goes to infinity with high probability and the probability of selection error degrades. As we can see, the FWER decreases at least exponentially with the number of samples. This behavior is consistent with Prop. 3.

*d) Application to Experimental Data.:* We illustrate the application of the proposed two-stage predictor on the Predictive Health and Disease dataset, which consists of gene expression levels and symptom scores of 38 different subjects. The data was collected during a challenge study for which some subjects become symptomatically ill with the H3N2 flu virus (Huang et al., 2011). For each subject, the gene expression levels and the symptoms have been recorded at a large number of time points that include pre-inoculation and post-inoculation sample times. 10 different symptom scores were measured. Each symptom score takes an integer value from 0 to 4, which measures the severity of that symptom at the corresponding time. The goal here is to learn a predictor that can accurately predict the symptom scores of a subject based on his measured gene expression levels.

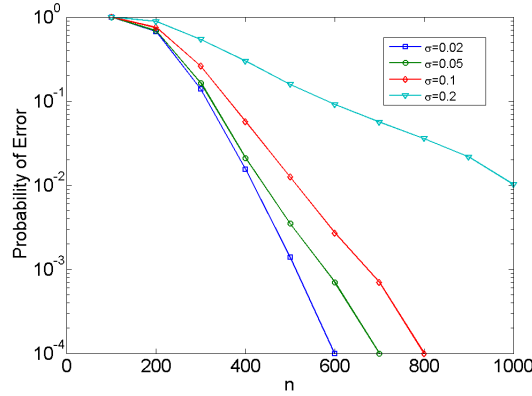


Fig. 9. Probability of selection error as a function of number of samples for PCS. The entries of the coefficient matrix are i.i.d. draws from distribution (29).

The number of predictor variables (genes) selected in the first stage is restricted to 50. Since, the symptom scores take integer values, the second stage uses multinomial logistic regression instead of the OLS predictor. The performance is evaluated by leave-one-out cross validation. To do this, the data from all except one subject are used as training samples and the data from the remaining subject are used as the test samples. The final MSE is then computed as the average over the 38 different leave-one-out cross validation trials. In each of the experiments 18 out of the 37 subjects of the training set, are used in first stage and all of the 37 subjects are used in the second stage. It is notable that except for the first two symptoms, PCS performs better in predicting the symptom scores.

Note that, in this experiment, each symptom is considered as a one dimensional response and the two-stage algorithm is applied to each symptom separately.

## VIII. CONCLUSION

We proposed an algorithm called Predictive Correlation Screening (PCS) for approximating the p-values of candidate predictor variables in high dimensional linear regression under a sparse null hypothesis. Variable selection was then performed based on the approximated p-values. PCS is specifically useful in cases where  $n \ll p$  and the high cost of assaying all regressor variables justifies a two-stage design: high throughput variable selection followed by predictor construction using fewer selected variables. Asymptotic analysis and experiments showed advantages of PCS as compared to LASSO and correlation learning.



Symptom	MSE: LASSO	MSE: PCS
Runny Nose	0.3346	0.3537
Stuffy Nose	0.5145	0.5812
Sneezing	0.4946	0.3662
Sore Throat	0.3602	0.3026
Earache	0.0890	0.0761
Malaise	0.4840	0.3977
Cough	0.2793	0.2150
Shortness of Breath	0.1630	0.1074
Headache	0.3966	0.3299
Myalgia	0.3663	0.3060
Average for all symptoms	0.3482	0.3036

TABLE I

MSE OF THE TWO-STAGE LASSO PREDICTOR AND THE PROPOSED TWO-STAGE PCS PREDICTOR USED FOR SYMPTOM SCORE PREDICTION. THE DATA COME FROM A CHALLENGE STUDY EXPERIMENT THAT COLLECTED GENE EXPRESSION AND SYMPTOM DATA FROM HUMAN SUBJECTS (HUANG ET AL., 2011).

## REFERENCES

- Arratia, R., Goldstein, L., and Gordon, L. Poisson approximation and the chen-stein method. *Statistical Science*, 5(4):403–424, 1990.
- Arya, S., Mount, D.M., Netanyahu, N.S., Silverman, R., and Wu, A.Y. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *Journal of the ACM (JACM)*, 45(6):891–923, 1998.
- Audibert, Jean-Yves, Munos, Rémi, and Szepesvári, Csaba. Tuning bandit algorithms in stochastic environments. In *Algorithmic Learning Theory*, pp. 150–165. Springer, 2007.
- Buehlmann, P. Boosting for high-dimensional linear models. *The Annals of Statistics*, 34(2):559–583, 2006.
- Bühlmann, P. and Van De Geer, S. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.
- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.
- Fan, Jianqing and Lv, Jinchi. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.
- Friedman, J., Hastie, T., and Tibshirani, R. *The elements of statistical learning*, volume 1. Springer Series in Statistics, 2001.
- Genovese, Christopher R, Jin, Jiashun, Wasserman, Larry, and Yao, Zhigang. A comparison of the lasso and marginal regression. *The Journal of Machine Learning Research*, 98888:2107–2143, 2012.
- Guyon, I. and Elisseeff, A. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.
- Haupt, Jarvis, Castro, Rui M, and Nowak, Robert. Distilled sensing: Adaptive sampling for sparse detection and estimation. *Information Theory, IEEE Transactions on*, 57(9):6222–6235, 2011.
- Hero, A. and Rajaratnam, B. Large-scale correlation screening. *Journal of the American Statistical Association*, 106(496):1540–1552, 2011.
- Hero, A. and Rajaratnam, B. Hub discovery in partial correlation graphs. *Information Theory, IEEE Transactions on*, 58(9):6064–6078, 2012.
- Huang, Yongsheng, Zaas, Aimee K, Rao, Arvind, Dobigeon, Nicolas, Woolf, Peter J, Veldman, Timothy, Øien, N Christine, McClain, Micah T, Varkey, Jay B, Nicholson, Bradley, et al. Temporal dynamics of host molecular responses differentiate symptomatic and asymptomatic influenza a infection. *PLoS*

- genetics*, 7(8):e1002234, 2011.
- Jégou, H., Douze, M., and Schmid, C. Product quantization for nearest neighbor search. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(1):117–128, 2011.
- Kim, J. and Park, H. Fast active-set-type algorithms for  $l_1$ -regularized linear regression. *Proc. AISTAT*, pp. 397–404, 2010.
- Obozinski, G., Wainwright, M.J., and Jordan, M.I. High-dimensional union support recovery in multivariate regression. *Advances in Neural Information Processing Systems*, 21, 2008.
- Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.
- Yuan, M. and Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2005.

## IX. APPENDIX

*Proof of Prop. 1:*

Define  $\phi_i^x = I(d_i^x \geq \delta)$ , where  $d_i^x$  is the degree of vertex  $i$  in part  $x$  in the thresholded correlation graph. We have:  $N_{\delta, \rho}^{xy} = \sum_{i=1}^p \phi_i^x$ . Define  $\phi_{ij}^{xy} = I(\mathbf{U}_j^y \in A(r, \mathbf{U}_i^x))$ , where  $A(r, \mathbf{U}_i^x)$  is the union of two anti-polar caps in  $S_{n-2}$  of radius  $\sqrt{2(1-\rho)}$  centered at  $\mathbf{U}_i^x$  and  $-\mathbf{U}_i^x$ .  $\phi_i^x$  can be expressed as:

$$\phi_i^x = \sum_{l=\delta}^q \sum_{\vec{k} \in \check{C}(q,l)} \prod_{j=1}^l \phi_{ik_j}^{xy} \prod_{m=l+1}^q (1 - \phi_{ik_m}^{xy}), \quad (30)$$

where  $\vec{k} = (k_1, \dots, k_q)$  and  $\check{C}(q, l) = \{\vec{k} : k_1 < k_2 < \dots < k_l, k_{l+1} < \dots < k_q, k_j \in \{1, 2, \dots, q\}, k_i \neq k_j\}$ .

By subtracting  $\sum_{\vec{k} \in \check{C}(q,l)} \prod_{j=1}^{\delta} \phi_{ik_j}^{xy}$  from both sides, we get:

$$\begin{aligned} \phi_i^x - \sum_{\vec{k} \in \check{C}(q,l)} \prod_{j=1}^{\delta} \phi_{ik_j}^{xy} = \\ \sum_{l=\delta+1}^q \sum_{\vec{k} \in \check{C}(q,l)} \prod_{j=1}^l \phi_{ik_j}^{xy} \prod_{m=l+1}^q (1 - \phi_{ik_m}^{xy}) + \\ \sum_{\vec{k} \in \check{C}(q,\delta)} \sum_{m=\delta+1}^q (-1)^{m-\delta} \prod_{j=1}^{\delta} \phi_{ik_j}^{xy} \\ \sum_{k'_{\delta+1} < \dots < k'_m, \{k'_{\delta+1}, \dots, k'_m\} \subset \{k_{\delta+1}, \dots, k_q\}} \prod_{n=\delta+1}^m \phi_{ik'_n}^{xy}. \end{aligned} \quad (31)$$

The following inequality will be helpful:

$$E[\prod_{i=1}^k \phi_{i,j_i}^{xy}] = \int_{S_{n-2}} dv \int_{A(r,v)} du_1 \dots \int_{A(r,v)} du_k f_{U_{i_1}^y, \dots, U_{i_k}^y, U_i^x}(u_1, \dots, u_k, v) \quad (32)$$

$$\leq P_0^k a_n^k M_{K|1}^{yx}, \quad (33)$$

where  $M_{K|1}^{yx} = \max_{i_1 \neq \dots \neq i_k, i} \|f_{U_{i_1}^y, \dots, U_{i_k}^y, U_i^x}\|_{\infty}$ .

Also we have:

$$E[\prod_{l=1}^m \phi_{i_l, j_l}^{xy}] \leq P_0^m a_n^m M_{|Q|}^{yx}, \quad (34)$$

where  $Q = \text{unique}(\{i_l, j_l\})$  is the set of unique indices among the distinct pairs  $\{\{i_l, j_l\}\}_{l=1}^m$  and  $M_{|Q|}^{yx}$  is a bound on the joint density of  $\mathbf{U}_Q^{xy}$ .

Now define:

$$\theta_i^x = \binom{q}{\delta}^{-1} \sum_{\vec{k} \in \check{C}(q,\delta)} \prod_{j=1}^{\delta} \phi_{ik_j}^{xy}. \quad (35)$$

Now, we show that

$$|E[\phi_i^x] - \binom{q}{\delta} E[\theta_i^x]| \leq \gamma_{q,\delta} (qP_0)^{\delta+1}, \quad (36)$$

where  $\gamma_{q,\delta} = 2e \max_{\delta+1 \leq l \leq q} \{a_n^l M_{l|1}^{yx}\}$ . To show this, take expectations from both sides of equation (31) and apply the bound (33) to obtain:

$$\begin{aligned} & |E[\phi_i^x] - \binom{q}{\delta} E[\theta_i^x]| \\ & \leq \sum_{l=\delta+1}^q \binom{q}{l} P_0^l a_n^l M_{l|1}^{yx} + \\ & \quad \binom{q}{\delta} \sum_{l=1}^{q-\delta} \binom{q-\delta}{l} P_0^{\delta+l} a_n^{\delta+l} M_{\delta+l|1}^{yx} \\ & \leq \max_{\delta+1 \leq l \leq q} \{a_n^l M_{l|1}^{yx}\} \\ & \quad \left( \sum_{l=\delta+1}^q \binom{q}{l} P_0^l + \binom{q}{\delta} P_0^\delta \sum_{l=1}^{q-\delta} \binom{q-\delta}{l} P_0^l \right) \\ & \leq \max_{\delta+1 \leq l \leq q} \{a_n^l M_{l|1}^{yx}\} \\ & \quad \left( (e - \sum_{l=1}^{\delta} \frac{1}{l!}) (qP_0)^{\delta+1} + \frac{q^\delta}{\delta!} P_0^\delta (e-1)(q-\delta)P_0 \right) \\ & \leq \max_{\delta+1 \leq l \leq q} \{a_n^l M_{l|1}^{yx}\} 2e (qP_0)^{\delta+1}, \end{aligned} \quad (37)$$

in which, the third inequality follows from the assumption  $qP_0 \leq 1$  along with the inequality :

$$\begin{aligned} & \sum_{k=s+1}^G \binom{G}{k} \left(\frac{t}{G}\right)^k \leq \sum_{k=s+1}^G \frac{t^k}{k!} \\ & \leq (e - \sum_{k=0}^s \frac{1}{k!}) t^{s+1}, \quad 0 \leq t \leq 1. \end{aligned} \quad (38)$$

Application of the mean value theorem to the integral representation (32) yields:

$$|E[\theta_i^x] - P_0^\delta J(\overline{f_{\mathbf{U}_{*1}^y, \dots, \mathbf{U}_{*s}^y, \mathbf{U}_i^x}})| \leq \tilde{\gamma}_{q,\delta}^{yx} (qP_0)^\delta r, \quad (39)$$

where  $\tilde{\gamma}_{q,\delta}^{yx} = 2a_n^{\delta+1} \dot{M}_{\delta+1|1}^{yx} / \delta!$  and  $\dot{M}_{\delta+1|1}^{yx}$  is a bound on the norm of the gradient:

$$\nabla_{\mathbf{U}_{i1}^y, \dots, \mathbf{U}_{i\delta}^y} \overline{f_{\mathbf{U}_{*1}^y, \dots, \mathbf{U}_{*s}^y, \mathbf{U}_i^x}}(\mathbf{U}_{i1}^y, \dots, \mathbf{U}_{i\delta}^y | \mathbf{U}_i^x). \quad (40)$$

Combining (37) and (39) and using the relation  $r = O((1-\rho)^{1/2})$  we conclude:

$$\begin{aligned} & |E[\phi_i^x] - \binom{q}{\delta} P_0^\delta J(\overline{f_{\mathbf{U}_i^x, \mathbf{U}_{*1}^y, \dots, \mathbf{U}_{*s}^y}})| \leq \\ & \quad O(p^\delta (qP_0)^\delta \max\{pP_0, (1-\rho)^{1/2}\}). \end{aligned} \quad (41)$$

Summing up over  $i$  we conclude:

$$\begin{aligned} E[N_{\delta,\rho}^{xy}] - \xi_{p,q,n,\delta,\rho}^{xy} J(\overline{f_{\mathbf{U}_*^x, \mathbf{U}_{\bullet_1}^y, \dots, \mathbf{U}_{\bullet_\delta}^y}}) &\leq \\ O(p(pP_0)^\delta \max\{pP_0, (1-\rho)^{1/2}\}) & \\ = O((\eta_{p,q,\delta}^{xy})^\delta \max\{\eta_{p,q,\delta}^{xy} p^{-\frac{1}{\delta}}, (1-\rho)^{1/2}\}), & \end{aligned} \quad (42)$$

where  $\eta_{p,q,\delta}^{xy} = p^{1/\delta} q P_0$ . This concludes (16).

To prove the second part of the theorem, we use Chen-Stein method (Arratia et al., 1990). Define:

$$\tilde{N}_{\delta,\rho}^{xy} = \sum_{0 \leq i_0 \leq p, 0 \leq i_1 < \dots < i_\delta \leq q} \prod_{j=1}^{\delta} \phi_{i_0 i_j}^{xy}. \quad (43)$$

Assume the vertices  $i$  in part  $x$  and  $y$  of the thresholded graph are shown by  $i^x$  and  $i^y$  respectively. for  $\vec{i} = (i_0^x, i_1^y, \dots, i_\delta^y)$ , define the index set  $B_{\vec{i}}^{xy} = B_{(i_0^x, i_1^y, \dots, i_\delta^y)}^{xy} = \{(j_0^x, j_1^y, \dots, j_\delta^y) : j_1^x \in \mathcal{N}_k^{xy}(i_1^x) \cup i_1^x, j_l^y \in \mathcal{N}_k^{xy}(i_l^y) \cup i_l^y, l = 1, \dots, \delta\} \cap C_{<}^{xy}$  where  $C_{<}^{xy} = \{(j_0, \dots, j_\delta) : 1 \leq j_0 \leq p, 1 \leq j_1 < \dots < j_\delta \leq q\}$ . Note that  $|B_{\vec{i}}^{xy}| \leq k^{\delta+1}$ . We have:

$$\tilde{N}_{\delta,\rho}^{xy} = \sum_{\vec{i} \in C_{<}^{xy}} \prod_{j=1}^{\delta} \phi_{i_0 i_j}^{xy}. \quad (44)$$

Assume  $N_{\delta,\rho}^{*xy}$  is a Poisson random variable with  $E[N_{\delta,\rho}^{*xy}] = \tilde{N}_{\delta,\rho}^{xy}$ . Using theorem 1 of (Arratia et al., 1990), we have:

$$2 \max_A |p(\tilde{N}_{\delta,\rho}^{xy} \in A) - p(N_{\delta,\rho}^{*xy} \in A)| \leq b_1 + b_2 + b_3, \quad (45)$$

where:

$$b_1 = \sum_{\vec{i} \in C_{<}^{xy}} \sum_{\vec{j} \in B_{\vec{i}}^{xy} - \vec{i}} E[\prod_{l=1}^{\delta} \phi_{i_0 i_l}^{xy}] E[\prod_{m=1}^{\delta} \phi_{j_0 j_m}^{xy}], \quad (46)$$

$$b_2 = \sum_{\vec{i} \in C_{<}^{xy}} \sum_{\vec{j} \in B_{\vec{i}}^{xy} - \vec{i}} E[\prod_{l=1}^{\delta} \phi_{i_0 i_l}^{xy} \prod_{m=1}^{\delta} \phi_{j_0 j_m}^{xy}], \quad (47)$$

and for  $p_{\vec{i}^{xy}} = E[\prod_{l=1}^{\delta} \phi_{i_0 i_l}^{xy}]$ :

$$b_3 = \sum_{\vec{i} \in C_{<}^{xy}} E[E[\prod_{l=1}^{\delta} \phi_{i_0 i_l}^{xy} - p_{\vec{i}^{xy}} | \vec{j}^x : \vec{j} \notin B_{\vec{i}}^{xy}]]. \quad (48)$$

Using the bound (34),  $E[\prod_{l=1}^{\delta} \phi_{i_0 i_l}^{xy}]$  is of order  $O(P_0^\delta)$ . Therefore:

$$\begin{aligned} b_1 &\leq O(pq^\delta k^{\delta+1} P_0^{2\delta}) = \\ &= O((\eta_{p,q,\delta}^{xy})^{2\delta} (k/(p^{\frac{1}{\delta+1}} q^{\frac{\delta}{\delta+1}}))^{\delta+1}). \end{aligned} \quad (49)$$

Note that, since  $\vec{i} \neq \vec{j}$ ,  $\prod_{l=1}^{\delta} \phi_{i_0 i_l}^{xy} \prod_{m=1}^{\delta} \phi_{j_0 j_m}^{xy}$  is a multiplication of at least  $\delta + 1$  different characteristic functions. Hence by (34),

$$E\left[\prod_{l=1}^{\delta} \phi_{i_0 i_l}^{xy} \prod_{m=1}^{\delta} \phi_{j_0 j_m}^{xy}\right] = O(P_0^{\delta+1}). \quad (50)$$

Hence,  $b_2 \leq O(pq^{\delta} k^{\delta+1} P_0^{\delta+1}) = O((\eta_{p,q,\delta}^{xy})^{\delta+1} (k/(p^{\frac{1}{\delta}} q)^{1/(\delta+1)})^{\delta+1})$ . Finally, to bound  $b_3$  we have:

$$\begin{aligned} b_3 &= \sum_{\vec{i} \in C_{<}^{xy}} E[E[\prod_{l=1}^{\delta} \phi_{i_0 i_l}^{xy} - p_{\vec{i}^{xy}} | \mathbf{U}_{A_k^{xy}(\vec{i})}]] = \\ &= \sum_{\vec{i} \in C_{<}^{xy}} \int_{S_{n-2}^{|A_k^{xy}(\vec{i})|}} dz_{i_0^x} \left( \prod_{l=1}^{\delta} \int_{S_{n-2}} dz_{i_l^x} \int_{A(r, \mathbf{u}_{i_0^x}^x)} d\mathbf{u}_{i_l^y}^y \right) \\ &\quad \left( \frac{f_{\mathbf{U}_{\vec{i}}^{xy} | \mathbf{U}_{A_k^{xy}(\vec{i})}}(\mathbf{U}_{\vec{i}}^{xy} | \mathbf{U}_{A_k^{xy}(\vec{i})}) - f_{\mathbf{U}_{\vec{i}}^{xy}}(\mathbf{U}_{\vec{i}}^{xy})}{f_{\mathbf{U}_{\vec{i}}^{xy}}(\mathbf{U}_{\vec{i}}^{xy})} \right) \\ &\quad f_{\mathbf{U}_{\vec{i}}^{xy}}(\mathbf{U}_{\vec{i}}^{xy}) f_{\mathbf{U}_{A_k^{xy}(\vec{i})}}(\mathbf{u}_{A_k^{xy}(\vec{i})}^x) \\ &\leq O(pq^{\delta} P_0^{\delta+1} \|\Delta_{p,q,n,k,\delta}^{xy}\|_1) = O((\eta_{p,q,\delta}^{xy})^{\delta} \|\Delta_{p,q,n,k,\delta}^{xy}\|_1). \end{aligned} \quad (52)$$

Therefore:

$$\begin{aligned} &|p(N_{\delta,\rho}^{xy} > 0) - (1 - \exp(-\Lambda_{\delta}^{xy}))| \leq \\ &|p(N_{\delta,\rho}^{xy} > 0) - (\tilde{N}_{\delta,\rho}^{xy} > 0)| + \\ &|p(\tilde{N}_{\delta,\rho}^{xy} > 0) - (1 - \exp(-E[\tilde{N}_{\delta,\rho}^{xy}]))| + \\ &|\exp(-E[\tilde{N}_{\delta,\rho}^{xy}]) - \exp(-\Lambda_{\delta}^{xy})| \\ &\leq 0 + b_1 + b_2 + b_3 + O(|E[\tilde{N}_{\delta,\rho}^{xy}] - \Lambda_{\delta}^{xy}|). \end{aligned} \quad (53)$$

Hence, it remains to bound  $O(|E[\tilde{N}_{\delta,\rho}^{xy}] - \Lambda_{\delta}^{xy}|)$ . Application of mean value theorem to the multiple integral (32) gives:

$$|E[\prod_{l=1}^{\delta} \phi_{i_l i_l}^{xy}] - P_0^{\delta} J(f_{\mathbf{U}_{i_1}^y, \dots, \mathbf{U}_{i_{\delta}}^y, \mathbf{U}_{i_0}^x})| \leq O(P_0^{\delta} r). \quad (54)$$

Using relation (44) we conclude:

$$\begin{aligned} &|E[\tilde{N}_{\delta,\rho}^{xy}] - p\binom{q}{\delta} P_0^{\delta} J(\overline{f_{\mathbf{U}_{*1}^y, \dots, \mathbf{U}_{*_{\delta}}^y, \mathbf{U}_{*0}^x}})| \leq \\ &O(pq^{\delta} P_0^{\delta} r) = O((\eta_{p,q,\delta}^{xy})^{\delta} r). \end{aligned} \quad (55)$$

Combining this with inequality (53) along with the bounds on  $b_1, b_2$  and  $b_3$ , completes the proof of (17).  $\square$

*Proof of Prop. 2:*

We prove the more general proposition below. Prop. 2 is then a direct consequence.

*Proposition:* Let  $\mathbb{X}$  and  $\mathbb{Y}$  be  $n \times p$  and  $n \times q$  data matrices whose rows are i.i.d. realizations of elliptically distributed  $p$ -dimensional and  $q$ -dimensional vectors  $\mathbf{X}$  and  $\mathbf{Y}$  with mean parameters  $\mu_x$  and  $\mu_y$  and covariance parameters  $\Sigma_x$  and  $\Sigma_y$ , respectively and cross covariance  $\Sigma_{xy}$ . Let  $\mathbb{U}^x = [\mathbf{U}_1^x, \dots, \mathbf{U}_p^x]$  and  $\mathbb{U}^y = [\mathbf{U}_1^y, \dots, \mathbf{U}_q^y]$  be the matrices of correlation  $U$ -scores. Assume that the covariance matrices  $\Sigma_x$  and  $\Sigma_y$  are block-sparse of degrees  $d_x$  and  $d_y$ , respectively (i.e. by rearranging their rows and columns, all non-diagonal entries are zero except a  $d_x \times d_x$  or a  $d_y \times d_y$  block). Assume also that the cross covariance matrix  $\Sigma^{xy}$  is block-sparse of degree  $d_1$  for  $x$  and degree  $d_2$  for  $y$  (i.e. by rearranging its rows and columns, all entries are zero except a  $d_1 \times d_2$  block), then

$$\tilde{\mathbb{U}}^x = \mathbb{U}^x(1 + O(d_x/p)). \quad (56)$$

Also assume that for  $\delta \geq 1$  the joint density of any distinct set of  $U$ -scores  $\mathbf{U}_i^x, \mathbf{U}_{i_1}^y, \dots, \mathbf{U}_{i_\delta}^y$  is bounded and differentiable over  $S_{n-2}^{\delta+1}$ . Then the  $(\delta + 1)$ -fold average function  $J(\overline{f_{\mathbf{U}_\bullet^x, \mathbf{U}_{*1}^y, \dots, \mathbf{U}_{*_\delta}^y}})$  and the average dependency coefficient  $\|\Delta_{p,n,k,\delta}^{xy}\|$  satisfy

$$J(\overline{f_{\mathbf{U}_\bullet^x, \mathbf{U}_{*1}^y, \dots, \mathbf{U}_{*_\delta}^y}}) = 1 + O(\max\{\frac{d_1}{p}, \delta \frac{(d_y - 1)}{q}\}), \quad (57)$$

$$\|\Delta_{p,q,n,k,\delta}^{xy}\|_1 = 0. \quad (58)$$

Furthermore,

$$J(\overline{f_{\tilde{\mathbf{U}}_\bullet^x, \mathbf{U}_{*1}^y, \dots, \mathbf{U}_{*_\delta}^y}}) = 1 + O(\max\{\frac{d_x}{p}, \frac{d_1}{p}, \delta \frac{(d_y - 1)}{q}\}) \quad (59)$$

$$\|\Delta_{p,q,n,k,\delta}^{\tilde{xy}}\|_1 = O((d_x/p)). \quad (60)$$

*Proof:* We have:

$$\tilde{\mathbb{U}}^x = (\mathbb{U}^x(\mathbb{U}^x)^T)^{-1} \mathbb{U}^x \mathbf{D}_{(\mathbb{U}^x)^T(\mathbb{U}^x(\mathbb{U}^x)^T)^{-2} \mathbb{U}^x}^{-\frac{1}{2}}. \quad (61)$$

By block sparsity of  $\Sigma_x, \mathbb{U}^x$  can be partitioned as:

$$\mathbb{U}^x = [\underline{\mathbb{U}}^x, \overline{\mathbb{U}}^x], \quad (62)$$

where  $\underline{\mathbb{U}}^x = [\mathbf{U}_1^x, \dots, \mathbf{U}_{d_x}^x]$  and  $\overline{\mathbb{U}}^x = [\mathbf{U}_{d_x+1}^x, \dots, \mathbf{U}_{p-d_x}^x]$  are dependent and independent columns of  $\mathbb{U}^x$ , respectively. Similarly, by block sparsity of  $\Sigma_y$ ,

$$\mathbb{U}^y = [\underline{\mathbb{U}}^y, \overline{\mathbb{U}}^y], \quad (63)$$



where  $\underline{\mathbb{U}}^y = [\underline{\mathbb{U}}_1^y, \dots, \underline{\mathbb{U}}_{d_y}^y]$  and  $\overline{\mathbb{U}}^y = [\overline{\mathbb{U}}_1^y, \dots, \overline{\mathbb{U}}_{q-d_y}^y]$  are dependent and independent columns of  $\mathbb{U}^y$ , respectively. By block sparsity of  $\Sigma_{xy}$ , at most  $d_1$  variables among  $\overline{\mathbb{U}}_1^x, \dots, \overline{\mathbb{U}}_{p-d_x}^x$  are correlated with columns of  $\mathbb{U}^y$ . Assume the correlated variables are among  $\overline{\mathbb{U}}_1^x, \dots, \overline{\mathbb{U}}_{d_2}^x$ . Similarly, at most  $d_2$  variables among  $\overline{\mathbb{U}}_1^y, \dots, \overline{\mathbb{U}}_{q-d_y}^y$  are correlated with columns of  $\mathbb{U}^x$ . Without loss of generality, assume the correlated variables are among  $\overline{\mathbb{U}}_1^y, \dots, \overline{\mathbb{U}}_{d_1}^y$ .

The columns of  $\overline{\mathbb{U}}^x$ , are i.i.d. and uniform over the unit sphere  $S_{n-2}$ . Therefore, as  $p \rightarrow \infty$ :

$$\frac{1}{p-d_x} \overline{\mathbb{U}}^x (\overline{\mathbb{U}}^x)^T \rightarrow E[\overline{\mathbb{U}}_1^x (\overline{\mathbb{U}}_1^x)^T] = \frac{1}{n-1} \mathbf{I}_{n-1}. \quad (64)$$

Also, since the entries of  $1/d_x \underline{\mathbb{U}}^x (\underline{\mathbb{U}}^x)^T$  are bounded by one, we have:

$$\frac{1}{p} \underline{\mathbb{U}}^x (\underline{\mathbb{U}}^x)^T = \mathbf{O}(d_x/p), \quad (65)$$

where  $\mathbf{O}(u)$  is an  $(n-1) \times (n-1)$  matrix whose entries are  $O(u)$ . Hence:

$$\begin{aligned} (\mathbb{U}^x (\mathbb{U}^x)^T)^{-1} \mathbb{U}^x &= \underline{\mathbb{U}}^x (\underline{\mathbb{U}}^x)^T + \overline{\mathbb{U}}^x (\overline{\mathbb{U}}^x)^T \mathbb{U}^x \\ &= \frac{n-1}{p} (\mathbf{I}_{n-1} + \mathbf{O}(d_x/p))^{-1} \mathbb{U}^x \\ &= \frac{n-1}{p} \mathbb{U}^x (1 + O(d_x/p)). \end{aligned} \quad (66)$$

Hence, as  $p \rightarrow \infty$ :

$$\begin{aligned} (\mathbb{U}^x)^T (\mathbb{U}^x (\mathbb{U}^x)^T)^{-2} \mathbb{U}^x &= \\ &= \left(\frac{n-1}{p}\right)^2 (\mathbb{U}^x)^T \mathbb{U}^x (1 + O(d_x/p)). \end{aligned} \quad (67)$$

Thus:

$$\mathbf{D}_{(\mathbb{U}^x)^T (\mathbb{U}^x (\mathbb{U}^x)^T)^{-2} \mathbb{U}^x} = \left( \frac{p}{n-1} \mathbf{I}_{n-1} (1 + O(d_x/p)) \right). \quad (68)$$

Combining (68) and (66) concludes (56).

Now we prove relations (57) and (58). Define the partition  $\mathcal{C} = \mathcal{D} \cup \mathcal{D}^c$  of the index set  $\mathcal{C}$  defined in (12), where  $\mathcal{D} = \{\vec{i} = (i_0, i_1, \dots, i_\delta) : i_0 \text{ is among } p-d_1 \text{ columns of } \mathbb{U}^x \text{ that are uncorrelated of columns of } \mathbb{U}^y \text{ and at most one of } i_1, \dots, i_\delta \text{ is less than or equal to } d_y\}$  is the set of  $(\delta+1)$ -tuples restricted to columns of  $\mathbb{U}^x$  and  $\mathbb{U}^y$  that are independent. We have:

$$\begin{aligned} J(\overline{f_{\mathbf{U}_{\bullet}^x, \mathbf{U}_{s_1}^y, \dots, \mathbf{U}_{s_\delta}^y}}) &= |\mathcal{C}|^{-1} 2^{-\delta} \sum_{s_1, \dots, s_\delta \in \{-1, 1\}} \\ & \left( \sum_{\vec{i} \in \mathcal{D}} + \sum_{\vec{i} \in \mathcal{D}^c} \right) J(f_{s_0 \mathbf{U}_{i_0}^x, s_1 \mathbf{U}_{i_1}^y, \dots, s_\delta \mathbf{U}_{i_\delta}^y}), \end{aligned} \quad (69)$$

and

$$\|\Delta_{p,q,n,k,\delta}^{xy}\|_1 = |\mathcal{C}|^{-1} \left( \sum_{\vec{i} \in \mathcal{D}} + \sum_{\vec{i} \in \mathcal{D}^c} \right) \Delta_{p,q,n,k,\delta}^{xy}(\vec{i}). \quad (70)$$

But,  $J(f_{s_0 \mathbf{U}_{i_0}^x, s_1 \mathbf{U}_{i_1}^y, \dots, s_\delta \mathbf{U}_{i_\delta}^y}) = 1$  for  $\vec{i} \in \mathcal{D}$  and  $\Delta_{p,q,n,k,\delta}^{xy}(\vec{i}) = 0$  for  $\vec{i} \in \mathcal{C}$ . Moreover, we have:

$$\frac{|\mathcal{D}|}{|\mathcal{C}|} = O\left(\frac{(p-d_1)(q-d_y+1)^\delta}{pq^\delta}\right). \quad (71)$$

Thus:

$$J(\overline{f_{\mathbf{U}_{\bullet}^x, \mathbf{U}_{*1}^y, \dots, \mathbf{U}_{*_\delta}^y}}) = 1 + O\left(\max\left\{\frac{d_1}{p}, \delta \frac{(d_y-1)}{q}\right\}\right). \quad (72)$$

Moreover, since  $\tilde{\mathbb{U}}^x = \mathbb{U}^x(1 + O(d_x/p))$ ,  $f_{\tilde{\mathbf{U}}_{i_0}^x, \mathbf{U}_{i_1}^y, \dots, \mathbf{U}_{i_\delta}^y} = f_{\mathbf{U}_{i_0}^x, \mathbf{U}_{i_1}^y, \dots, \mathbf{U}_{i_\delta}^y}(1 + O(d_x/p))$ . This concludes:

$$J(\overline{f_{\tilde{\mathbf{U}}_{\bullet}^x, \mathbf{U}_{*1}^y, \dots, \mathbf{U}_{*_\delta}^y}}) = 1 + O\left(\max\left\{\frac{d_x}{p}, \frac{d_1}{p}, \delta \frac{(d_y-1)}{q}\right\}\right), \quad (73)$$

and

$$\|\Delta_{p,q,n,k,\delta}^{\tilde{xy}}\|_1 = O(d_x/p). \quad (74)$$

□

*Proof of Proposition 3:* First we prove the theorem for  $q = 1$ . Without loss of generality assume

$$Y = a_1 X_1 + a_2 X_2 + \dots + a_k X_k + \sigma N, \quad (75)$$

where  $N$  follows the standard normal distribution. Note that since  $q = 1$ ,  $a_1, \dots, a_k$  are scalars.

Defining  $\mathbf{b} = \Sigma^{1/2} \mathbf{a}$ , the response  $Y$  can be written as:

$$Y = a_1 Z_1 + a_2 Z_2 + \dots + a_k Z_k + \sigma N, \quad (76)$$

in which  $Z_1, \dots, Z_k$  are i.i.d. standard normal random variables. Assume  $\mathbf{U}_1, \dots, \mathbf{U}_p, \mathbf{U}_N$  represent the U-scores (which are in  $S_{n-2}$ ) corresponding to  $Z_1, \dots, Z_p, N$ , respectively. It is easy to see:

$$\mathbf{U}_y = \frac{b_1 \mathbf{U}_1 + b_2 \mathbf{U}_2 + \dots + b_k \mathbf{U}_k + \sigma \mathbf{U}_N}{\|b_1 \mathbf{U}_1 + b_2 \mathbf{U}_2 + \dots + b_k \mathbf{U}_k + \sigma \mathbf{U}_N\|}. \quad (77)$$

If  $\mathbf{U}$  and  $\mathbf{V}$  are the U-scores corresponding to two random variables, and  $r$  is the correlation coefficient between the two random variables, we have:

$$|r| = 1 - \frac{(\min\{\|\mathbf{U} - \mathbf{V}\|, \|\mathbf{U} + \mathbf{V}\|\})^2}{2}. \quad (78)$$

Let  $r_{y,i}$  represent the sample correlation between  $Y$  and  $X_i$ . Here, we want to upper bound  $\text{prob}\{|r_{y,1}| < |r_{y,k+1}|\}$ . We have:

$$\begin{aligned} \text{prob}\{|r_{y,1}| < |r_{y,k+1}|\} &= \\ \text{prob}\{1 - \frac{(\min\{\|\mathbf{U}_1 - \mathbf{U}_y\|, \|\mathbf{U}_1 + \mathbf{U}_y\|\})^2}{2} &< \\ 1 - \frac{(\min\{\|\mathbf{U}_{k+1} - \mathbf{U}_y\|, \|\mathbf{U}_{k+1} + \mathbf{U}_y\|\})^2}{2}\} &= \end{aligned} \quad (79)$$

$$\begin{aligned} \text{prob}\{\min\{\|\mathbf{U}_1 - \mathbf{U}_y\|, \|\mathbf{U}_1 + \mathbf{U}_y\|\} &> \\ \min\{\|\mathbf{U}_{k+1} - \mathbf{U}_y\|, \|\mathbf{U}_{k+1} + \mathbf{U}_y\|\}\} &\leq \end{aligned} \quad (80)$$

$$\begin{aligned} \text{prob}\{\|\mathbf{U}_1 - \mathbf{U}_y\| &> \\ \min\{\|\mathbf{U}_{k+1} - \mathbf{U}_y\|, \|\mathbf{U}_{k+1} + \mathbf{U}_y\|\}\} &= \\ \text{prob}\{\{\|\mathbf{U}_1 - \mathbf{U}_y\| > \|\mathbf{U}_{k+1} - \mathbf{U}_y\|\} \cup \\ \{\|\mathbf{U}_1 - \mathbf{U}_y\| > \|\mathbf{U}_{k+1} + \mathbf{U}_y\|\}\} &\leq \\ \text{prob}\{\|\mathbf{U}_1 - \mathbf{U}_y\| > \|\mathbf{U}_{k+1} - \mathbf{U}_y\|\} &+ \\ \text{prob}\{\|\mathbf{U}_1 - \mathbf{U}_y\| > \|\mathbf{U}_{k+1} + \mathbf{U}_y\|\} &= \end{aligned} \quad (81)$$

$$2 \text{ prob}\{\|\mathbf{U}_1 - \mathbf{U}_y\| > \|\mathbf{U}_{k+1} - \mathbf{U}_y\|\}, \quad (82)$$

in which, the last inequality holds since  $\mathbf{U}_{k+1}$  is uniform over  $S_{n-2}$  and is independent of  $\mathbf{U}_1$  and  $\mathbf{U}_y$ . Therefore, it suffices to upper bound  $p_1 := \text{prob}\{\|\mathbf{U}_1 - \mathbf{U}_y\| > \|\mathbf{U}_{k+1} - \mathbf{U}_y\|\}$ . Define:

$$\mathbf{V} = b_2 \mathbf{U}_2 + \cdots + b_k \mathbf{U}_k, \quad (83)$$

and

$$\mathbf{U}_* = \mathbf{V} / \|\mathbf{V}\|. \quad (84)$$

By symmetry,  $\mathbf{U}_*$  is uniform over  $S_{n-2}$ . Hence:

$$\mathbf{U}_y = \frac{b_1 \mathbf{U}_1 + \|\mathbf{V}\| \mathbf{U}_*}{\|b_1 \mathbf{U}_1 + \|\mathbf{V}\| \mathbf{U}_*\|}. \quad (85)$$

Since  $\|\mathbf{V}\| \leq |b_2| + \cdots + |b_k|$ , we have:

$$\frac{|b_1|}{\|\mathbf{V}\|} \geq \frac{|b_1|}{|b_2| + \cdots + |b_k|} = \frac{|b_1|}{c_1}, \quad (86)$$

where  $c_1 := |b_2| + \cdots + |b_k|$ . Define:

$$\theta_1 = \cos^{-1}(\mathbf{U}_y^T \mathbf{U}_1), \quad (87)$$

and

$$\theta_1 = \cos^{-1}(\mathbf{U}_y^T \mathbf{U}_*). \quad (88)$$

It is easy to see that:

$$\frac{\sin \theta_1}{\sin \theta_2} \leq \frac{c_1}{|b_1|}. \quad (89)$$

For each  $0 \leq \theta \leq \pi$ , define:

$$\beta_1(\theta) = \max_{0 \leq \theta' \leq \pi} \frac{\theta}{\theta + \theta'} \quad \text{s.t.} \quad \frac{\sin \theta}{\sin \theta'} \leq \frac{c_1}{|b_1|}. \quad (90)$$

Now fix the point  $\mathbf{U}_1$  on  $S_{n-2}$ . Define  $f(\theta)$  as the probability distribution of  $\theta_2$ . Also, define  $p(\theta)$  as the probability that the angle between the uniformly distributed (over  $S_{n-2}$ ) point  $\mathbf{U}_{k+1}$  and  $\mathbf{U}_y$  is less than  $\theta$ . Since  $\mathbf{U}_1$  is independent of  $\mathbf{U}_*$  and  $\mathbf{U}_{k+1}$  is independent of  $\mathbf{U}_y$ , clearly:

$$p(\theta) = \int_0^\theta f(\theta') d\theta'. \quad (91)$$

We have:

$$\begin{aligned} p_1 &\leq \int_0^\pi p(\beta_1(\theta)\theta) f(\theta) d\theta \\ &= \int_0^{\pi/2} (p(\beta_1(\theta)\theta) + p(\beta_1(\pi - \theta)(\pi - \theta))) f(\theta) d\theta, \end{aligned} \quad (92)$$

where the last equality holds because  $f(\theta) = f(\pi - \theta)$ . Noting the fact that:

$$\begin{aligned} \int_0^\pi p(\theta) f(\theta) d\theta &= \int_0^{\pi/2} (p(\theta) + p(\pi - \theta)) f(\theta) d\theta \\ &= \int_0^{\pi/2} f(\theta) d\theta = \frac{1}{2}. \end{aligned} \quad (93)$$

we conclude:

$$\begin{aligned} p_1 &\leq \frac{1}{2} - \int_0^{\pi/2} \{ (p(\theta) - p(\beta_1(\theta)\theta)) + \\ &\quad (p(\pi - \theta) - p(\beta_1(\pi - \theta)(\pi - \theta))) \} f(\theta) d\theta. \end{aligned} \quad (94)$$

Hence by (91), for any  $0 < \theta_0 < \pi/2$ :

$$p_1 \leq \frac{1}{2} - \int_{\theta_0}^{\pi/2} p_{\gamma_1}(\theta) f(\theta) d\theta, \quad (95)$$

in which

$$\begin{aligned} p_{\gamma_1}(\theta) &= p(\theta + \gamma_1\theta) - p(\theta - \gamma_1\theta) \\ &= \text{prob}\{\theta - \gamma_1\theta \leq \theta_2 \leq \theta + \gamma_1\theta\}, \end{aligned} \quad (96)$$

with

$$\gamma_1 = \min_{\theta_0 \leq \theta \leq \pi - \theta_0} 1 - \beta_1(\theta) = 1 - \max_{\theta_0 \leq \theta \leq \pi - \theta_0} \beta_1(\theta). \quad (97)$$

It is easy to check that  $\gamma_1 > 0$ . Therefore, since  $p_{\gamma_1}(\theta)$  is an increasing functions of  $\theta$  for  $0 \leq \theta \leq \pi/2$ , we conclude:

$$p_1 \leq \frac{1}{2} - \int_{\theta_0}^{\pi/2} p_{\gamma_1}(\theta_0) f(\theta) d\theta. \quad (98)$$

Choose  $\theta_0$  so that  $\theta_0 = \frac{\pi}{2+\gamma_1}$ . We have:

$$\begin{aligned} p_1 &\leq \frac{1}{2} - p_{\gamma_1}(\pi/(2+\gamma_1)) \int_{\pi/(2+\gamma_1)}^{\pi/2} f(\theta) d\theta \\ &= \frac{1}{2} - \int_{\pi(1-\gamma_1)/(2+\gamma_1)}^{\pi(1+\gamma_1)/(2+\gamma_1)} f(\theta) d\theta \int_{\pi/(2+\gamma_1)}^{\pi/2} f(\theta) d\theta \\ &\leq \frac{1}{2} - \int_{\pi/2-\gamma_1\pi/6}^{\pi/2+\gamma_1\pi/6} f(\theta) d\theta \int_{\pi/2-\gamma_1\pi/6}^{\pi/2} f(\theta) d\theta \\ &\leq \frac{1}{2} - 2 \left( \int_{\pi/2-\gamma_1\pi/6}^{\pi/2} f(\theta) d\theta \right)^2, \end{aligned} \quad (99)$$

in which, the last inequality holds, since  $0 < \gamma_1 < 1$ . Defining  $\lambda_1 = \sin(\pi/2 - \gamma_1\pi/6)$  and using the formula for the area of the spherical cap, we will have:

$$\begin{aligned} \int_{\pi/2-\gamma_1\pi/6}^{\pi/2} f(\theta) d\theta &= \\ \frac{I_1((n-2)/2, 1/2) - I_{\lambda_1}((n-2)/2, 1/2)}{2I_1((n-2)/2, 1/2)}, \end{aligned} \quad (100)$$

in which

$$I_x(a, b) = \frac{\int_0^x t^{a-1} (1-t)^{b-1} dt}{\int_0^1 t^{a-1} (1-t)^{b-1} dt}, \quad (101)$$

is the regularized incomplete beta function. Hence:

$$\int_{\pi/2-\gamma_1\pi/6}^{\pi/2} f(\theta) d\theta = \frac{\int_{\lambda_1}^1 t^{(n-4)/2} / \sqrt{1-t} dt}{2 \int_0^1 t^{(n-4)/2} / \sqrt{1-t} dt}. \quad (102)$$

Note that we have:

$$\begin{aligned} &\frac{\int_{\lambda_1}^1 t^{(n-4)/2} / \sqrt{1-t} dt}{2 \int_0^{\lambda_1} t^{(n-4)/2} / \sqrt{1-t} dt} \\ &\geq \frac{\int_{\lambda_1}^1 t^{(n-4)/2} / \sqrt{1-\lambda_1} dt}{2 \int_0^1 t^{(n-4)/2} / \sqrt{1-\lambda_1} dt} \\ &= \frac{1 - \lambda_1^{(n-2)/2}}{\lambda_1^{(n-2)/2}} := \kappa_1. \end{aligned} \quad (103)$$

Hence:

$$\begin{aligned} \int_{\pi/2-\gamma_1\pi/6}^{\pi/2} f(\theta)d\theta &\geq \frac{\int_{\lambda_1}^1 t^{(n-4)/2}/\sqrt{1-t}dt}{2(1+1/\kappa_1) \int_{\lambda_1}^1 t^{(n-4)/2}/\sqrt{1-t}dt} \\ &= \frac{\kappa_1}{2(\kappa_1+1)} = \frac{1-\lambda_1^{(n-2)/2}}{2}. \end{aligned} \quad (104)$$

Hence by (99):

$$p_1 \leq \lambda_1^{(n-2)/2} - \lambda_1^{n-2} \leq \lambda_1^{(n-2)/2}. \quad (105)$$

Therefore,  $p_1$  decreases at least exponentially by  $n$ .

Assume  $P(i)$  for  $1 \leq i \leq k$ , represents the probability that the active variable  $X_i$  is not among the selected  $k$  variables. By (82) and using the union bound we have:

$$P(1) \leq 2(p-k)\lambda_1^{(n-2)/2}. \quad (106)$$

Similar inequalities can be obtained for  $P(2), \dots, P(k)$  which depend on  $\lambda_2, \dots, \lambda_k$ , respectively. Finally, using the union bound, the probability  $P$  that all the active variables are correctly selected satisfies:

$$P \geq 1 - 2(p-k) \sum_{i=1}^k \lambda_i^{(n-2)/2} \geq 1 - 2k(p-k)\lambda^{(n-2)/2}, \quad (107)$$

where  $\lambda := \max_{1 \leq i \leq k} \lambda_i$ . This concludes that if  $n = \Theta(\log p)$ , with probability at least  $1 - 1/p$  the exact support can be recovered using PCS.

For  $q > 1$ , by union bound, the probability of error becomes at most  $q$  times larger and this concludes the statement of proposition 3.  $\square$

*Proof of Proposition 4:* First we consider a two-stage predictor similar to the one introduced in previous section with the difference that the  $n$  samples which are used in stage 1 are not used in stage 2. Therefore, there are  $n$  and  $t-n$  samples used in the first and the second stages, respectively. Following the notation introduced in previous section, we represent this two-stage predictor by  $n|(t-n)$ . The asymptotic results for the  $n|(t-n)$  two-stage predictor will be shown to hold as well for the  $n|t$  two-stage predictor.

Using inequalities of the form (106) and the union bound, it is straightforward to see that for any subset  $\pi \neq \pi_0$  of  $k$  elements of  $\{1, \dots, p\}$ , the probability that  $\pi$  is the outcome of variable selection via PCS, is bounded above by  $2k(p-k)c_\pi^n$ , in which  $0 < c_\pi < 1$  is a constant that depends on the quantity

$$\min_{j \in \pi_0 \cap \pi^c} \frac{|a_j|}{\sum_{l \in \pi_0} |a_l|}. \quad (108)$$

The expected MSE of the  $n|(t-n)$  algorithm can be written as:

$$\mathbb{E}[\text{MSE}] = \sum_{\pi \in S_k^p, \pi \neq \pi_0} p(\pi) \mathbb{E}[\text{MSE}_\pi] + p(\pi_0) \mathbb{E}[\text{MSE}_{\pi_0}], \quad (109)$$

where  $S_k^p$  is the set of all  $k$ -subsets of  $\{1, \dots, p\}$ ,  $p(\pi)$  is the probability that the outcome of variable selection via PCS is the subset  $\pi$ , and  $\text{MSE}_\pi$  is the MSE of OLS stage when the indices of the selected variables are the elements of  $\pi$ . Therefore using the bound (107), the expected MSE is upper bounded as below:

$$\begin{aligned} \mathbb{E}[\text{MSE}] &\leq 2k(p-k) \sum_{\pi \in S_k^p, \pi \neq \pi_0} c_\pi^n \mathbb{E}[\text{MSE}_\pi] + \\ &\quad (1 - 2k(p-k)c_0^n) \mathbb{E}[\text{MSE}_{\pi_0}], \end{aligned} \quad (110)$$

$c_0$  is a constant that depends on the quantity (26). It can be shown that if there is at least one wrong variable selected ( $\pi \neq \pi_0$ ), the OLS estimator is biased and the expected MSE converges to a positive constant  $M_\pi$  as  $(t-n) \rightarrow \infty$ . When all the variables are selected correctly (subset  $\pi_0$ ), MSE goes to zero with rate  $O(1/(t-n))$ . Hence:

$$\begin{aligned} \mathbb{E}[\text{MSE}] &\leq 2k(p-k) \sum_{\pi \in S_k^p, \pi \neq \pi_0} c_\pi^n M_\pi + \\ &\quad (1 - 2k(p-k)c_0^n) O(1/(t-n)) \leq \\ &\quad 2k(p-k)C_1 C^n + (1 - 2k(p-k)C^n)C_2/(t-n), \end{aligned} \quad (111)$$

where  $C, C_1$  and  $C_2$  are constants that do not depend on  $n$  or  $p$  but depend on the quantities  $\sum_{j \in \pi_0} a_j^2$  and  $\min_{j \in \pi_0} |a_j| / \sum_{l \in \pi_0} |a_l|$ .

On the other hand since at most  $t$  variables could be used in OLS stage, the expected MSE is lower bounded:

$$\mathbb{E}[\text{MSE}] \geq \Theta(1/t). \quad (112)$$

It can be seen that the minimum of (111) as a function of  $n$ , subject to the constraint (23), happens for  $n = O(\log t)$  if  $\Theta(\log t) \leq \frac{\mu - tk}{p-k}$ ; otherwise it happens for 0. If  $\Theta(\log t) \leq \frac{\mu - tk}{p-k}$ , the minimum value attained by the upper bound (111) is  $\Theta(1/t)$  which is as low as the lower bound (112). This shows that for large  $t$ , the optimal number of samples that should be assigned to the PCS stage of the  $n|(t-n)$  predictor is  $n = O(\log t)$ . As  $t \rightarrow \infty$ , since  $n = O(\log t)$ , the MSE of the  $n|t$  predictor proposed in Sec. V converges to the MSE of the  $n|(t-n)$  predictor. Therefore, as  $t \rightarrow \infty$ ,  $n = O(\log t)$  becomes optimal for the  $n|t$  predictor as well.  $\square$